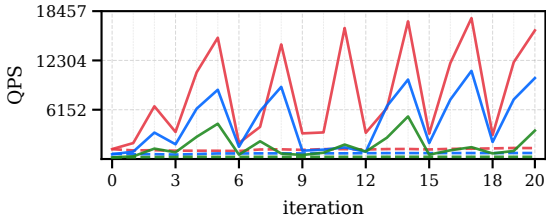


Query Throughput



- | | | |
|------------------------|-------------------------|-------------------------|
| -- Backend only (k=1) | -- Backend only (k=100) | -- With QVCache (k=10) |
| -- Backend only (k=10) | -- With QVCache (k=1) | -- With QVCache (k=100) |